Frequentist Evaluation of Intervals Estimated for a Binomial Parameter and for the Ratio of Poisson Means

Robert D. Cousins, Kathryn E. Hymes, Jordan Tucker

Dept. of Physics and Astronomy, University of California, Los Angeles, California 90095, USA

Abstract

Confidence intervals for a binomial parameter or for the ratio of Poisson means are commonly desired in high energy physics (HEP) applications such as measuring a detection efficiency or branching ratio. Due to the discreteness of the data, in both of these problems the frequentist coverage probability unfortunately depends on the unknown parameter. Trade-offs among desiderata have led to numerous sets of intervals in the statistics literature, while in HEP one typically encounters only the classic intervals of Clopper-Pearson (central intervals with no undercoverage but substantial over-coverage) or a few approximate methods which perform rather poorly. If strict coverage is relaxed, some sort of averaging is needed to compare intervals. In most of the statistics literature, this averaging is over different values of the unknown parameter, which is conceptually problematic from the frequentist point of view in which the unknown parameter is typically fixed. In contrast, we perform an (unconditional) *average over observed data* in the ratio-of-Poisson-means problem. If strict conditional coverage is desired, we recommend Clopper-Pearson intervals and intervals from inverting the likelihood ratio test (for central and non-central intervals, respectively). Lancaster's mid-P modification to either provides excellent unconditional average coverage in the ratio-of-Poisson-means problem.

Key words: hypothesis test, confidence interval, binomial parameter, ratio of Poisson means
PACS: 06.20.Dk, 07.05.Kf

Email addresses: cousins@physics.ucla.edu (Robert D. Cousins), tucker@physics.ucla.edu (Jordan Tucker).

1 Introduction

The construction of confidence intervals for a binomial parameter (probability of success in a binomial distribution), while already performed by Clopper and Pearson (C-P) in 1934 [1], remains a topic of discussion in the modern statistics literature due to differences in opinion about the best way to deal with imperfect coverage rooted in the discreteness of the observed number of successes. Clopper and Pearson's central intervals, while guaranteeing no undercoverage, result in considerable overcoverage (conservatism). Numerous alternatives have been put forward in the intervening years, with reviews such as that by Brown, Cai, and Dasgupta [2] recommending for general use some sets of intervals which are less conservative than those of C-P, but which undercover for certain values of the binomial parameter. In this paper, we examine the problem from the point of view of high energy physics (HEP) applications, including the problem of confidence intervals for the ratio of Poisson means. The latter problem provides an additional frequentist criterion, not yet considered by Brown et al., for judging the merits of sets of intervals for a binomial parameter.

Figure 1a illustrates the issue to be addressed. For each value of the binomial parameter ρ, one supposes that it is the true but unknown value, and calculates the long-run fraction of experiments for which that value is contained in ("covered by") the reported confidence intervals. In Fig. 1a, the number of trials is fixed at 10, the probabilities for the number of successes are calculated from the binomial formula using the true value of ρ, and the central C-P confidence intervals with a confidence level (C.L.) of 68.27% are used. The upper and lower endpoints of the C-P interval are, respectively, 15.87% C.L. lower and upper one-sided confidence limits. The coverage of the one-sided confidence limits is always greater than or equal to 15.87%, with equality on a discrete finite set of values. As this set of points is different for upper and lower confidence limits, the coverage of the two-sided intervals in this example is always strictly greater than 68.27%, an unfortunate consequence of the discrete nature of the observation.

For comparison, Fig. 1b is the coverage plot for central intervals derived using a Bayesian technique with Jeffreys prior, as described below. The coverage oscillates around the nominal 68.27%, in a way that by eye seems to have an "average" value near 68.27%. The problem from the frequentist point of view is that such averaging over values of the unknown parameter is typically not appropriate since the unknown true value of ρ is fixed, i.e., not sampled from a distribution.

The effect of discreteness is also displayed in Figs. 2a and b, which show the coverage as a function of n_{tot}, for fixed $\rho = 0.1$. The above four plots are

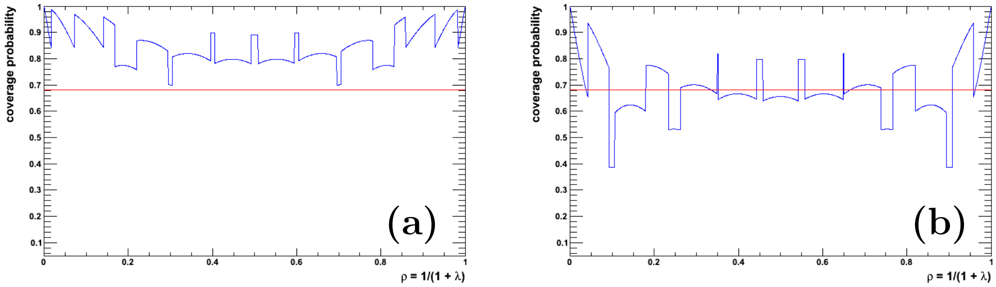

Fig. 1. (a) Coverage of 68.27% C.L. Clopper-Pearson intervals, and (b) coverage of intervals calculated using a Bayesian method with Jeffreys prior and containing 68.27% posterior probability, both as a function of ρ, for fixed $n_{\mathrm{tot}} = 10$. (a) and (b) are horizontal slices of Figs. 3a and b, respectively.

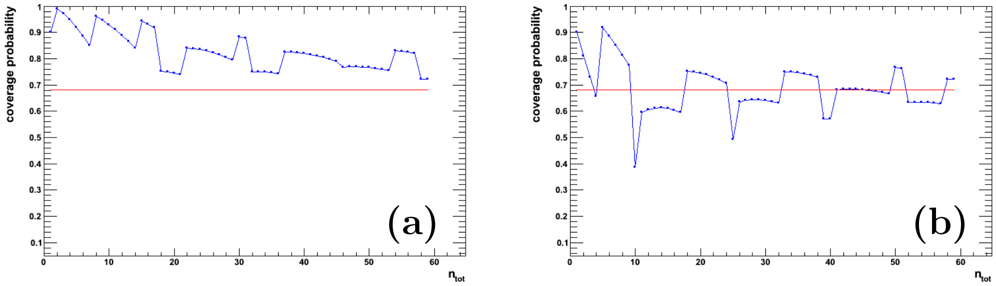

Fig. 2. (a) Coverage of 68.27% C.L. Clopper-Pearson intervals, and (b) coverage of intervals calculated using a Bayesian method with Jeffreys prior and containing 68.27% posterior probability, as a function of n_{tot}, for fixed $\rho = 0.1$. (a) and (b) are vertical slices of Figs. 3a and b, respectively.

horizontal and vertical slices of a much larger pattern of behavior displayed in Figs 3a and b. In these two figures, and corresponding figures below, ΔCL is the difference between the actual coverage and the nominal coverage, in this case 68.27%.

These are but two of many sets of intervals that have been proposed. The saw-tooth features of the coverage plots are intrinsic to all methods except the randomization technique (mentioned in Sec. 2) which brings other difficulties. Which sets are deemed preferable depends on the value one attaches to never having undercoverage, on what sort of averaging (if any) over values of ρ one allows, whether or not one desires central intervals, and additional issues such as whether one is especially concerned about behavior near the endpoints, $\rho = 0$ and 1.

In this paper, we emphasize that a *frequentist* averaging method, which averages over repeatedly sampled *data*, can be used to evaluate sets of intervals, in contrast to most previous averaging methods which average over the parameter ρ in some metric. The frequentist average is performed by using the strong connection between confidence intervals for a binomial parameter and

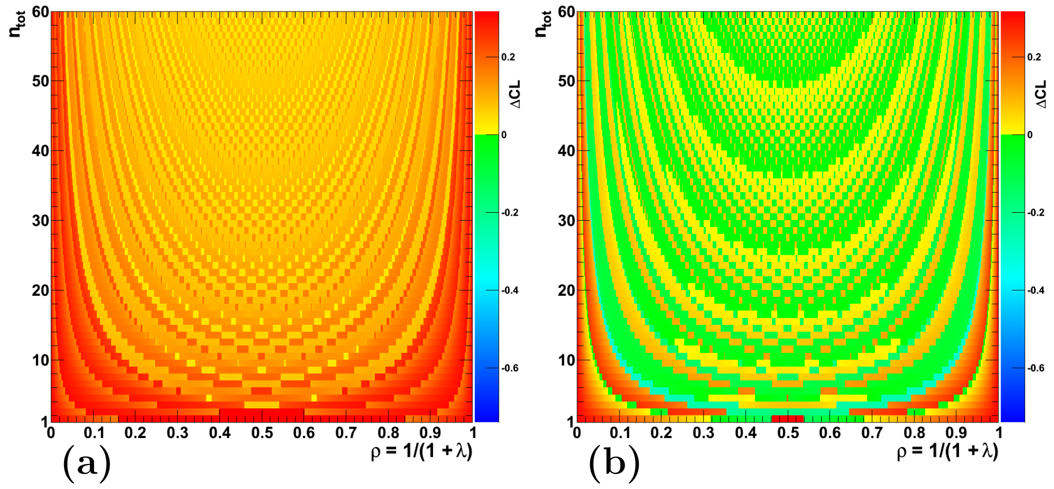

Fig. 3. (a) Coverage of 68.27% C.L. Clopper-Pearson intervals, and (b) coverage of intervals calculated using a Bayesian method with Jeffreys prior and containing 68.27% posterior probability, as a function of ρ and n_{tot}. ΔCL is the difference between the actual coverage and the nominal coverage, 68.27%.

confidence intervals for the ratio of two unknown Poisson means. For pairs of integers sampled from two fixed but unknown Poisson means, fluctuations in the total number of observed events provides a random sampling which partially smoothes out the saw-tooth structure seen in binomial coverage plots. Said another way (using terminology defined below) we use the unconditional global coverage as a criterion for averaging over imperfect conditional coverage of each fixed total number of events.

In the traditional definition of "confidence interval", defined by Neyman as we discuss below, the name implies no undercoverage for any value of the unknown parameter. When dealing with approximate methods, immaterial departures from perfect coverage are typically tolerated as long as it is clearly understood that coverage is only approximate. When methods yield intervals which are known to have non-negligible undercoverage for some values of the unknown parameter (such as for the mid-P intervals for the binomial parameter), the statistics literature is mixed on whether or not to refer to these intervals as confidence intervals. In this paper, we attempt to follow HEP practice by requiring no undercoverage when referring to intervals as "confidence intervals".

In Sec. 2, we review the relevant concepts from interval and hypothesis test construction and define the notation. In Sec. 3, we briefly describe a number of papers from the vast literature on binomial intervals. In Sec. 4, we generalize to the ratio-of-Poisson-means problem, and review some relevant literature. In Sec. 5 we present our results on the coverage of a number of the methods. We conclude in Sec. 6.

2 Definitions and Notation

We let $\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}},\rho)$ denote the probability of n_{on} successes in n_{tot} trials, each with binomial parameter ρ:

$$\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}},\rho) = \frac{n_{\mathrm{tot}}!}{n_{\mathrm{on}}!(n_{\mathrm{tot}}-n_{\mathrm{on}})!}\,\rho^{n_{\mathrm{on}}}\,(1-\rho)^{(n_{\mathrm{tot}}-n_{\mathrm{on}})}. \tag{1}$$

In repeated trials, n_{on} has mean

$$n_{\mathrm{tot}}\rho \tag{2}$$

and rms deviation

$$\sqrt{n_{\mathrm{tot}}\rho(1-\rho)}. \tag{3}$$

For asymptotically large n_{tot}, Bi can be approximated by a normal distribution with this mean and rms deviation.

With observed number of successes n_{on}, the likelihood function $\mathcal{L}(\rho)$ follows from reading Eqn. 1 as a function of ρ. The maximum is at

$$\hat{\rho} = n_{\mathrm{on}}/n_{\mathrm{tot}}. \tag{4}$$

In some applications, n_{tot} is not fixed but is itself a random variable sampled from a Poisson distribution with mean μ_{tot}:

$$\mathrm{Poi}(n_{\mathrm{tot}}|\mu_{\mathrm{tot}}) = \frac{\mathrm{e}^{-(\mu_{\mathrm{tot}})}\,(\mu_{\mathrm{tot}})^{n_{\mathrm{tot}}}}{n_{\mathrm{tot}}!}. \tag{5}$$

In this case, n_{on} and $n_{\mathrm{off}} = n_{\mathrm{tot}} - n_{\mathrm{on}}$ can be considered to be independent random variables, each sampled from a Poisson distribution with means μ_{on} and μ_{off}, respectively, satisfying

$$\mu_{\mathrm{on}} + \mu_{\mathrm{off}} = \mu_{\mathrm{tot}}. \tag{6}$$

The ratio of the Poisson means is then

$$\lambda = \mu_{\mathrm{off}}/\mu_{\mathrm{on}}, \tag{7}$$

and the binomial parameter can be written as

$$\rho = \mu_{\mathrm{on}}/\mu_{\mathrm{tot}} = 1/(1+\lambda). \tag{8}$$

The joint probability $P(n_{\rm on}, n_{\rm off})$ of observing $n_{\rm on}$ and $n_{\rm off}$ can then be expressed in two equivalent ways: as the product of independent Poisson probabilities for $n_{\rm on}$ and $n_{\rm off}$; or as the product of a single Poisson probability with mean $\mu_{\rm tot}$ for the total number of events $n_{\rm tot}$, and the binomial probability that this total is divided as observed:

$$\begin{aligned} P(n_{\rm on}, n_{\rm off}) &= \frac{{\rm e}^{-\mu_{\rm on}}\, \mu_{\rm on}^{n_{\rm on}}}{n_{\rm on}!} \times \frac{{\rm e}^{-\mu_{\rm off}}\, \mu_{\rm off}^{n_{\rm off}}}{n_{\rm off}!} \\ &= \frac{{\rm e}^{-(\mu_{\rm on}+\mu_{\rm off})}\, (\mu_{\rm on}+\mu_{\rm off})^{n_{\rm tot}}}{n_{\rm tot}!} \times \qquad (9) \\ & \frac{n_{\rm tot}!}{n_{\rm on}!(n_{\rm tot}-n_{\rm on})!}\, \rho^{n_{\rm on}}\, (1-\rho)^{(n_{\rm tot}-n_{\rm on})}. \qquad (10) \end{aligned}$$

In more compact notation, we have:

$$\begin{aligned} P(n_{\rm on}, n_{\rm off}) &= {\rm Poi}(n_{\rm on}|\mu_{\rm on})\ {\rm Poi}(n_{\rm off}|\mu_{\rm off}) \qquad (11) \\ &= {\rm Poi}(n_{\rm tot}|\mu_{\rm tot})\ {\rm Bi}(n_{\rm on}|n_{\rm tot}, \rho). \qquad (12) \end{aligned}$$

This observation is the basis of hypothesis tests on the ratio of Poisson means going back to Przyborowski and Wilenski [3] in 1940, as recommended in HEP by James and Roos [4], and as discussed by statistician Reid [5]. All the dependence on ratio of Poisson means λ is in the *conditional* binomial probability for the observed "successes" $n_{\rm on}$, *given* the observed total number of events $n_{\rm tot} = n_{\rm on} + n_{\rm off}$.

We consider a general parameter θ (such as ρ or λ) and randomly sampled data (such as $n_{\rm on}$ or other observables), the probability of which depends on θ. We then consider a recipe for computing the endpoints of a confidence interval $[t_{\rm low}, t_{\rm up}]$ for θ, as functions of the (randomly sampled) data. (In this paper we always include the endpoints in the confidence interval.) The set of all confidence intervals obtainable from all possible data sets using this recipe is called a *confidence set*. For each value of θ, one can then compute the probability that that θ is contained in ("covered by") the confidence intervals in the confidence set, for data sampled according to that θ. Normally it is highly desirable that this coverage probability be independent of θ, and is called the confidence coefficient or (more commonly in HEP) the confidence level (C.L.) of the confidence set. For situations such as those in this paper, in which the data takes on only discrete values, the coverage probability depends on θ, as illustrated above in Figs. 1a and b.

In classical hypothesis testing, a common hypothesis test is that which tests the hypothesis H_0 that θ is equal to a particular value, θ_0, against the alternative that $\theta \neq \theta_0$. One constructs recipes for accepting/rejecting H_0 based on the (randomly sampled) data, the probability for which depends on θ. One defines the significance level α of the test (also called size of the test) as the probability of rejecting H_0 if is true; again it is desirable that α is independent of θ. In the formal theory of Neyman-Pearson hypothesis testing, α is specified in advance; once data are obtained, the *p-value* is the smallest value of α for which H_0 would be rejected.

As discussed by Kendall and Stuart and successors [6], one can construct a hypothesis test at significance level α simply by using a confidence set with C.L. $= 1 - \alpha$ and accepting H_0 if θ_0 is contained in the confidence interval for θ based on the obtained data. One can equally well derive confidence sets from any given recipe for testing the hypothesis $\theta = \theta_0$, simply by including in the interval those values of θ_0 which would not be rejected by such a test. This way of constructing confidence intervals is referred to in the statistics literature as "inverting the hypothesis test". (An example now familiar in the HEP literature is the set of intervals advocated by Feldman and Cousins [7], which are constructed by inverting the likelihood ratio test of Ref. [6].) It can happen that the resulting "intervals" are not simply connected, in which case various adjustments are typically made, for example adding to the interval any interior points not initially part of it (thus adding to the over-coverage).

In this duality, confidence intervals formed by inverting a test with significance level α have coverage probability $= 1 - \alpha$ under H_0, i.e.,

$$P(\theta_0 \in [t_{\rm low}, t_{\rm up}]) = 1 - \alpha. \tag{13}$$

Central confidence intervals have the additional property that the intervals $[t_{\rm low}, t_{\rm max}]$ and $[t_{\rm min}, t_{\rm up}]$ each separately have coverage probability $= 1 - \alpha/2$, i.e.,

$$P(\theta_0 \in [t_{\rm low}, t_{\rm max}]) = P(\theta_0 \in [t_{\rm min}, t_{\rm up}]) = 1 - \alpha/2, \tag{14}$$

where $t_{\rm max}$ and $t_{\rm min}$ are the maximum and minimum values of θ defined in the model (e.g., $t_{\rm min} = 0$ and $t_{\rm max} = 1$ if θ is a binomial parameter). In this case, for example, $t_{\rm up}$ is often referred to as a $(1 - \alpha/2)$ C.L. upper confidence limit for θ.

If, due to the discreteness, the significance level can only be specified to be less than or equal to α, then the equal signs in Eqns. 13 and 14 become "$\geq$".

Without invoking a Gaussian approximation in the construction of an interval itself, it is often useful to make the correspondence with the number of Gaussian standard deviations having a *single*-tailed probability equal to $\alpha/2$. Thus, Z denotes the number of standard deviations away from the center of a Gaussian distribution, with a subscript representing the (one-tailed) tail probability beyond that Z.

$$Z_{\alpha/2} = \Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(\alpha/2) \tag{15}$$

where

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} \exp(-t^2/2)\, dt \ = \ \frac{1+\mathrm{erf}(Z/\sqrt{2})}{2}, \tag{16}$$

so that

$$Z = \sqrt{2}\,\mathrm{erf}^{-1}(1-\alpha). \tag{17}$$

E.g., $Z_{\alpha/2} = 1$ for $\alpha/2 = 0.159$, and $Z_{\alpha/2} = 1.64$ for $\alpha/2 = 0.05$.

3 Recipes for intervals for ρ

A plethora of recipes exists for intervals approximating confidence intervals for binomial parameter ρ. They correspond to various choices regarding:

- Whether or not the intervals are central intervals;
- Whether or not the intervals are derived from rigorously inverting a hypothesis test (in which case, which test?);
- Whether or not an asymptotic approximation is invoked;
- Whether or not Bayesian machinery is used to derive the intervals;
- Whether or not so-called "corrections" are made in an attempt to improve the coverage probability.

As emphasized by Cai [8], some methods with bad properties as one-sided intervals have good properties as two-sided intervals due to cancellations in coverage departures between the two tails.

3.1 Asymptotic approximations

We begin with one of the most popular methods, which is also one of the worst-performing if not *the* worst-performing of popular methods. We follow the literature in referring to this interval as the *Wald interval.* After estimating

$$\hat{\rho} = n_{\mathrm{on}}/n_{\mathrm{tot}}, \tag{18}$$

the Wald method invokes the Gaussian approximation *without properly inverting the hypothesis test against the null*, but rather simply substituting $\hat{\rho}$ for ρ into Eqn. 3 and using this *fixed* value of the rms to obtain the two endpoints,

$$\hat{\rho} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n_{\mathrm{tot}}}}. \tag{19}$$

Already in 1927, Edwin Wilson [9] realized that since the rms depends on the unknown parameter ρ, the more appropriate way to invoke the Gaussian approximation was by consistently inverting the test using the rms of the null hypothesis for each value of ρ. For the lower endpoint, one uses the lowest value ρ_1 such that $\rho_1 + Z_{\alpha/2}\sqrt{\rho_1(1-\rho_1)/n_{\mathrm{tot}}}$ contains $\hat{\rho}$. Analogously for the upper endpoint, one uses the largest value ρ_2 such that $\rho_2 - Z_{\alpha/2}\sqrt{\rho_2(1-\rho_2)/n_{\mathrm{tot}}}$ contains $\hat{\rho}$. Letting $T = (Z_{\alpha/2})^2/n_{\mathrm{tot}}$, this leads to a quadratic equation in ρ for the endpoints, $(\rho - \hat{\rho})^2 = T\rho(1-\rho)$, with solutions

$$\rho = \frac{\hat{\rho} + T/2}{1+T} \pm \frac{\sqrt{\hat{\rho}(1-\hat{\rho})T + T^2/4}}{1+T}. \tag{20}$$

These endpoints form the *Wilson score interval*; in spite of the fact that it is a non-iterative solution using nothing more than a square root, sadly it is commonly overlooked in favor of the Wald interval when a quick Gaussian estimate is desired.

Letting $\tilde{\rho}$ denote the midpoint of the Wilson score interval, from Eqn. 20 one has

$$\tilde{\rho} = \frac{\hat{\rho} + T/2}{1+T} = \frac{n_{\mathrm{on}} + (Z_{\alpha/2})^2/2}{n_{\mathrm{tot}} + (Z_{\alpha/2})^2}. \tag{21}$$

As discussed in detail by Agresti and Coull [10], $\tilde{\rho}$ differs from $\hat{\rho}$ by formally adding $(Z_{\alpha/2})^2$ to the number of actual trials, and making half of them successes. It thus "shrinks" the maximum-likelihood point estimate $\hat{\rho}$ towards 0.5. For 95% C.L., ($Z_{\alpha/2} = 1.96$), the easy-to-remember rule of thumb is simply "add four trials with two successes" to obtain the (approximate) Wilson midpoint. For quick estimates one can use $\tilde{\rho}$ rather than $\hat{\rho}$ (and $n_{\mathrm{tot}} + (Z_{\alpha/2})^2$ rather than n_{tot}) in the Wald formula (Eqn. 19) and obtain intervals with surprisingly decent coverage, much better than when using $\hat{\rho}$ (and avoiding the useless result at extreme data). We refer to such intervals as *generalized Agresti-Coull (AC) intervals* (adding "generalized" to the name given by Brown et al. [2] to distinguish from the simpler version). Agresti and Coull themselves (who regard the C-P intervals as not optimal for statistical practice due to their conservatism) advocate AC intervals for teaching and the Wilson score interval for statistical practice [11].

The asymptotic theory in which log likelihood-ratios are related to chi-square distributions [12,13] provides another interval estimate. In the present case, the interval consists of all points satisfying

$$Z^2_{\alpha/2} \ \geq \ -2\ln\frac{\mathcal{L}(\rho)}{\mathcal{L}(\hat{\rho})} \ = \ 2\ln\left(\frac{\hat{\rho}}{\rho}\right)^{n_{\rm on}} + 2\ln\left(\frac{1-\hat{\rho}}{1-\rho}\right)^{(n_{\rm tot}-n_{\rm on})} . \tag{22}$$

As there is more than one use of the likelihood ratio for intervals in this paper, we refer unambiguously to intervals from Eqn. 22 as $\Delta(-2\ln\mathcal{L})$ *intervals*. In addition to the usual caution required in using asymptotic formulas for small numbers of events, in the present case there are well-known issues at the extrema of ρ, where the conditions of the asymptotic theory justifying Eqn. 22 are not satisfied.

As discussed by Cox and Hinkley [14], for the exponential family of distributions, i.e., those of the form $p(\theta) = \exp(a(\theta)b(x) + c(\theta) + d(x))$, the transformation to the "natural parameter" $\phi = -a(\theta)$ and new data variable $z = b(x)$ leads to some mathematical simplifications. The natural parameter for the binomial distribution is the logit function,

$$\phi = \ln(\rho/(1-\rho)), \tag{23}$$

also known as the log odds ratio; it is a convenient map from (0,1) to $(-\infty,\infty)$ in a variety of contexts. Such non-linear maps are a reminder that the concept of "shortest" is metric-dependent: it is easy to find pairs of intervals whose relative length in ρ is reversed when transformed to ϕ.

The logit makes the mathematics simpler, but as Cox and Hinkley note, whether this is really the best parametrization can depend on other considerations as well. Models involving the logit and its inverse have a long history and were used in the work that was awarded the 2000 Nobel Memorial Prize in Economics. In any case, one can apply the same sort of Gaussian approximation to the logit ϕ as is applied in forming the Wald intervals for ρ. The maximum likelihood estimate $\hat{\phi}$ is obtained by plugging in $\hat{\rho}$. The variance of $\hat{\phi}$ is estimated as $n_{\rm tot}/(n_{\rm on}(n_{\rm tot} - n_{\rm on})) = 1/(\hat{\rho}(1-\hat{\rho}))$. One then has an interval for ϕ, which can be mapped into an interval for ρ. As the formulas are singular for $n_{\rm on} = 0$ and $n_{\rm on} = n_{\rm tot}$, patches are required, which are sometimes used for other values of $n_{\rm on}$ as well, in particular adding 1/2 to both numerator and denominator in the logit formula [2,15,16].

Given any test statistic and an ordering defined for it, confidence intervals with minimum guaranteed coverage can be constructed by the technique of Neyman [17], which corresponds to inverting a hypothesis test with rigorously calculated significance level. Such methods are often called "exact" since approximations are not made in the calculation of the probabilities, but as already shown for Clopper-Pearson, the coverage is by no means "exactly" equal to the nominal C.L.! Analogous to the Neyman construction described in detail for a similar discrete problem in Ref. [7], for each value of ρ one forms acceptance intervals by adding the probabilities $\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}}, \rho)$ for observed n_{on} until the threshold $1-\alpha$ is crossed. An auxiliary principle for the ordering in which the probabilities for the n_{on} are to be added to the acceptance set must be specified.

Clopper and Pearson [1] constructed central confidence intervals which remain the standard [18] for those who insist (as has been common in HEP) that coverage is always rigorously respected; the ordering is performed separately on each end of the acceptance interval. As noted above, the cost is severe over-coverage for some values of ρ. Angus and Schafer [19] compute over-coverage of C-P intervals, pointing out that $(1-\alpha)$ C.L. intervals can have coverage probability as high as $(1-\alpha/2)$ for some values of the true ρ; in fact the coverage is always this high or larger if n_{tot} is small enough that $n_{\mathrm{tot}} < (1 - \ln\alpha/\ln 2)$.

Sterne [20], followed soon by Crow [21], constructed sets of *non-central* intervals with guaranteed minimum coverage. The idea is to reduce over-coverage due to discreteness by relaxing the requirement in Eqn. 14 while retaining that in Eqn. 13. An obvious ordering principle to start with is based on $\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}}, \rho)$, i.e., adding points to (either end of) the acceptance interval in decreasing order of probability so as to minimize the length of the acceptance interval. There is room for adjustment, however, since in many cases the acceptance interval can be shifted, keeping its length fixed while still maintaining coverage. As there is considerable ambiguity in the best way to make such adjustments, there have been numerous attempts to improve upon Sterne's non-central intervals, variously referred to as *two-tailed* or *both-tailed* intervals.

Blyth and Still [22] give a very detailed discussion of the ambiguities encountered in such both-tailed constructions. They list some desirable features of intervals and, while giving their preferences for resolving ambiguities, note that "We see no way of combining these desirable properties into a precise criterion that would be generally accepted." Casella [23] reviews Blyth and Still and their predecessors and describes a method for systematically further reducing the length of confidence intervals obtained from such constructions: "...move all the lower endpoints of the intervals as far to the right as possible." In commenting [24] on Brown et al., he strongly advocates covering at the nominal value or greater, preferring the Blyth-Still intervals with his length-reduction algorithm. Blaker [25,26] discusses in enlightening detail various both-tailed methods, arriving at intervals which have good properties (nesting) when viewed as a function of the confidence level. But Vos and Hudson [27] explain in detail how both-tailed tests, even those of Blaker [25], inevitably have some undesirable behavior due to discreteness.

In HEP, Feldman and Cousins [7] popularized a Neyman construction that is equivalent to inverting the hypothesis test based on likelihood ratios. The likelihood-ratio ordering in the Neyman construction is based not on $\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}},\rho)$ as used by Sterne, but on the likelihood ratio $\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}},\rho)/\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}},\hat{\rho})$. The corresponding test, the likelihood ratio test (LR test), is one of the standard methods in classical statistics [6]. Coverage of both-tailed intervals for ρ from such "exact inversion of the LR test" was illustrated by statisticians Corcoran and Mehta [28], who prefer over-coverage to under-coverage, and who advocate either these intervals or the Blyth-Still-Casella intervals. Ranucci [29] compares coverage plots of intervals for ρ from such likelihood-ratio ordering with the intervals of C-P and of Sterne. As mentioned above (and described in Sec. IV of Ref. [7]), some interior points can be absent from the "interval" after first inverting the LR test; if so, in the present paper they are added in order to make the interval simply connected.

3.2.1 Randomization, Mid-P, and Continuity Correction

In order to remove the over-coverage in Neyman constructions caused by the discreteness of the integer-valued observations such as that of C-P, in 1950 Stevens [30] and others suggested adding a random number uniform on (0,1) to the observed integer, and performing the construction on the resultant continuous variable. As discussed in detail in Ref. [6], this technique, known as *randomization*, results in shorter intervals and perfect coverage. But as this extra random number was to be chosen from a table of (uniform) random numbers, it is rarely if ever used except in theoretical discussions. Reference [31] discusses how more meaningful data-based uniform variates can be justifiably used in randomization of Poisson observations, but we do not pursue this approach in this paper.

As an alternative to randomization, Lancaster [32] suggested in 1961 to deal with the discreteness issue in many distributions by quoting an intermediate value of the tail probability, since known as the "mid-P" value. For a one-sided test, it is the null probability of more extreme results *plus (only) half the probability of the observed* n_{on}. It corresponds to randomization always with the addition of 1/2 to the observed integer successes rather than addition of a uniform variate on (0,1). By using only half the probability rather than all the probability of the observed n_{on}, the mid-P is less than the strict p-value. As such, it has neither perfect coverage nor a guarantee against undercoverage, but the mid-P has attracted much more of a following than randomization, as the result is not influenced by an arbitrary random number. Berry and Armitage [33] review mid-P intervals in various contexts including the binomial problem, suggesting that they can be appropriate when combining results from several studies. Agresti and Gottard [34,35] further advocate mid-P intervals, provide a useful overview, and provide a function for computing them in the statistical package R.

Another commonly used device in dealing with discrete distributions is called (somewhat optimistically) a *continuity correction*, for example adding or subtracting 1/2 (or more generally another constant) from the observed number of successes. Although there is some advocacy of continuity corrections with respect to the binomial problem in the literature, it appears that there are better-performing ways to deal with the discreteness [2,36].

3.3 Bayesian-inspired methods

Intervals derived using Bayesian machinery [37] can be evaluated according to their frequentist coverage properties, and there has long been interest in prior probability density functions ("priors") which lead to Bayesian credible intervals possessing approximate frequentist coverage. Recent reviews of such "probability matching priors" are in Refs. [38,39]. Since the work of Welch and Peers [40,41,42], it has been recognized that Jeffreys's prior [43,44] (derived by Jeffreys under a different motivation) is the lowest-order probability matching prior for one parameter (although care must be taken in interpreting this result for a discrete distribution such as binomial). The Jeffreys prior for the binomial problem is

$$p(\rho) \propto \frac{1}{\sqrt{\rho(1-\rho)}}, \tag{24}$$

which is a special case (with $a = b = 1/2$) of the two-parameter *beta distribution*, which has pdf

$$p(\rho\,; a, b) \propto \rho^{a-1}(1-\rho)^{b-1}. \tag{25}$$

The beta distribution is closely linked to the binomial distribution [37], and varying a and b provides a family of priors (including the uniform prior with

$a = b = 1$). The posterior from a beta prior is also a beta distribution [37,45], and intervals can be obtained from it using various criteria such as length or centrality.

Geisser [46] considers several noninformative priors in the Bayesian literature and advocates a prior uniform in ρ, rejecting the Jeffreys prior because it violates the (strong) likelihood principle [37]. The Comments following Ref. [46] (by Bernardo, Novick, and Zellner) point out problems when ρ is near 0 or 1. Brenner and Quan [47] also advocate a prior uniform in ρ, apparently unaware of the many issues [44,46] in trying to represent "no prior information" in a prior. Copas [48] emphasizes that Bayesian-derived results, such as those of Brenner and Quan, do not automatically have good frequentist properties, and in particular criticize the prior uniform in ρ.

Rubin and Schenker [49] derive logit-based intervals using the Jeffreys prior, recalling earlier work (including Gart [15]) connecting this approach to using asymptotic logit estimation after appending a half success and a half failure as mentioned above. They calculate coverage both for fixed values of ρ and for values averaged over the Jeffreys prior.

3.4 Comparative studies

Given the abundance of methods, a number of authors have compared them by various criteria such as average coverage or average length (both of which are metric dependent), behavior near the extrema of ρ, etc. There is no general agreement on even basic features, such as whether or not coverage should be respected everywhere or in an average sense. And as noted above, preferences can differ if one is concerned only with one-sided intervals.

Reiczigel [50] advocates quoting an adjusted significance level based on calculated coverage rather than the nominal coverage used in the construction. Agresti [51] advocates inverting two-tailed tests (leading to non-central intervals) rather than two one-tailed tests. Puza and O'Neill [52] perform coverage studies and advocate a "new class" of C-P-inspired intervals which transition from one-sided to two-sided intervals.

Vollset [53] reviews in detail thirteen methods, recommending a continuity-corrected Wilson score method (strongly disfavored by Refs. [2,36]), but describing as "safe" the C-P intervals, mid-P intervals, and Wilson score intervals without the continuity correction. He finds likelihood-ratio intervals to be too narrow for boundary outcomes.

Edwardes [54] compared several methods using coverage averaged over a chosen metric, studying the behavior as a function of the constant used in the continuity correction. Among many results, he finds good performance for a Wald logit interval with *negative* continuity correction.

Newcombe [55] considers the "strict conservatism" of the C-P method to be "unnecessarily conservative" and compares it to the Wald and score methods, with and without continuity corrections, mid-P, and $\Delta(-2\ln\mathcal{L})$ methods, and appears to favor the mid-P and score methods.

Lu [56] compares lengths of intervals made with Bayesian methods with beta function priors, with endpoints adjusted according to Blyth [57]. He discusses in some detail the beta-distribution formulas and their numerical evaluation.

Agresti and Min [58] generally prefer both-tailed (non-central) tests if coverage is strictly required (unless one specifically requires a one-sided-test), and recommend mid-P tests if not. They also discuss using unconditional coverage in eliminating nuisance parameters in the context of the difference in binomial parameters.

Pires and Amado [59] compare 20 methods (counting various continuity corrections), with a table giving formulas for all of them. They prefer C-P intervals if coverage is strictly required, or the (ungeneralized) Agresti-Coull "add 4" method [10] if not.

Brown et al. [2] consider the Clopper-Pearson intervals to be "wastefully conservative", and advocate three sets of intervals with coverage oscillating about the nominal value: the Wilson score interval, the Agresti-Coull interval, and a Bayesian interval with Jeffreys prior and equal tails (except when n_{on} is at the extreme values, in which case they take one tail).

Coverage plots are illustrated in Refs. [2,28,29,35,48,49,50,51,52,53,54,58,59].

4 Application to the Ratio of Poisson Means

As discussed in Sec. 2, intervals for the ratio of Poisson means λ are readily obtained from intervals for the binomial parameter ρ, and vice versa. The *conditional coverage of* λ*, given* n_{tot}, can be read off coverage plots for ρ using $\lambda = (1-\rho)/\rho$. However, we can also consider the *unconditional coverage of* λ, as a function of the two unknown means μ_{on} and μ_{off}, as follows.

Given fixed μ_{on}, μ_{off} (and hence λ), a C.L., and a recipe for intervals, then for all pairs $(n_{\text{on}}, n_{\text{off}})$, one can calculate both the confidence interval for λ ($[t_{\text{low}}, t_{\text{up}}]$) and the probability $P(n_{\text{on}}, n_{\text{off}})$ of obtaining that pair (Eqn. 10). From these one can calculate probabilities that $\lambda < t_{\text{low}}$, that $\lambda \in [t_{\text{low}}, t_{\text{up}}]$ and that $\lambda > t_{\text{up}}$. Figures 4a and b have the results of such unconditional coverage

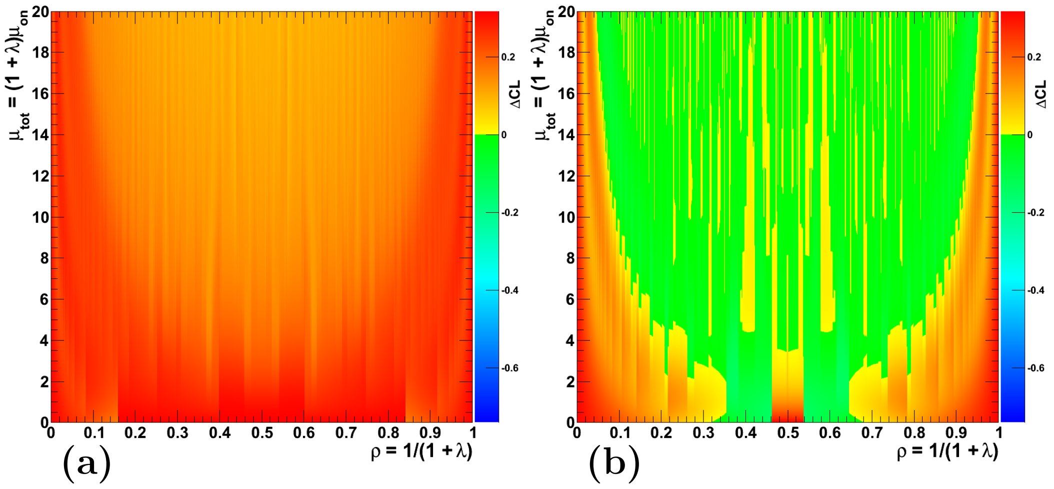

Fig. 4. Unconditional coverage of (a) 68.27% C.L. Clopper-Pearson intervals for the ratio of Poisson means λ and (b) intervals for λ calculated using a Bayesian method with Jeffreys prior and containing 68.27% posterior probability. As described in the text, the coverage as a function of $(\mu_{\rm on}, \mu_{\rm off})$ is displayed equivalently as a function of $(\rho, \mu_{\rm tot})$. ΔCL is the difference between the actual coverage and the nominal coverage, 68.27%.

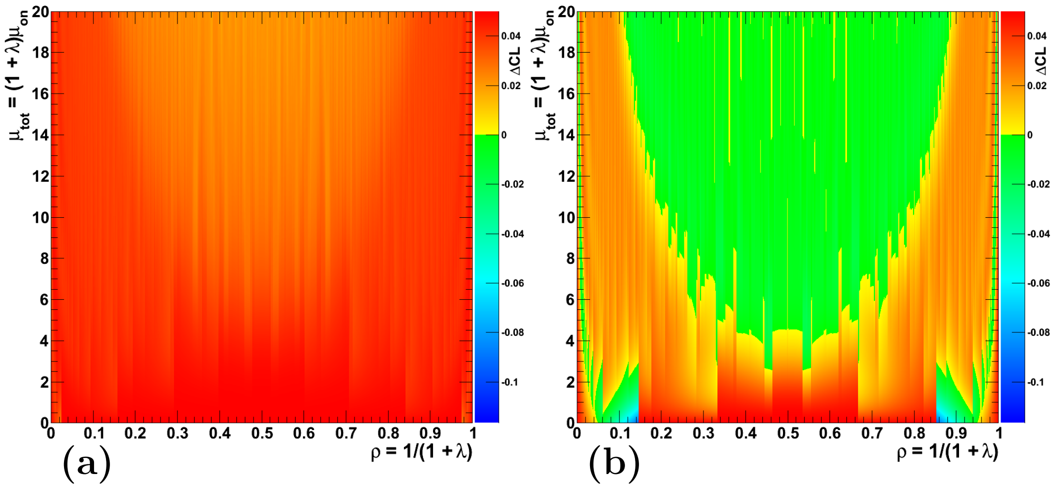

Fig. 5. Unconditional coverage of (a) 95% C.L. Clopper-Pearson intervals for λ and (b) intervals for λ calculated using a Bayesian method with Jeffreys prior and containing 95% posterior probability.

for the Clopper-Pearson and Jeffreys-prior-based recipes used in the previous figures; Figs. 5a and b contain the corresponding plots for 95% C.L. In order to facilitate comparison with the conditional coverage plots, the axes are $\mu_{\rm tot}$ and ρ, from which one can make the translation to $(\mu_{\rm on}, \mu_{\rm off})$ via $\mu_{\rm on} = \rho\mu_{\rm tot}$ and $\mu_{\rm off} = (1-\rho)\mu_{\rm tot}$.

These plots of unconditional coverage thus *average over observed* $n_{\rm tot}$ given true values of μ_{on} and μ_{off}, in contrast to nearly all previous studies which average over unknown true values of parameters. While the use of the unconditional ensemble (rather than the restricted "conditional" ensemble having the observed n_{tot}) goes against the mainstream statistical practice of using the conditional ensemble in a case such as this [5], we believe that this frequentist averaging over data provides at least as good a way to average-out some discreteness effects as does the common averaging over ρ, which requires a choice of metric (often ρ itself, although one can argue that the metric in which the prior is uniform is the natural metric in a Bayesian calculation). The issue is discussed in detail in Ref. [60], which as mentioned below describes a construction of central confidence intervals for the ratio of Poisson means having strict unconditional, but not conditional, coverage. (Averaging over observed *data* with different values of n_{tot} was used in Ref. [60] to cancel out some under- and over-coverage in different values of n_{tot} at each value of the ratio.)

As apparent from Fig. 4a, applying Clopper-Pearson binomial confidence intervals to the ratio-of-Poisson-means problem further propagates the over-coverage due to the discreteness. We return to this important point in Sec.5 below after first briefly reviewing some previous work applying non-C-P binomial intervals to the ratio-of-Poisson-means problem.

Price and Bonett [16] consider various solutions to the problem of the ratio of Poisson means from a broad point of view, including translating into this problem the binomial confidence intervals of C-P [1], Wilson score [9], and Agresti and Coull [10]. They also consider recipes derived directly for the ratio problem, namely a square-root transformation, an adjusted Wald log-linear model equivalent to the adjusted Wald logit formula mentioned above, and Bayesian methods including a Gamma prior for the ratio. Their conclusions depend as usual on considerations such as whether coverage is rigorously required, but tend to favor the adjusted Wald log-linear model in which 0.5 is added to the observed counts, resulting in endpoints

$$\frac{n_{\mathrm{on}}+0.5}{n_{\mathrm{tot}}-n_{\mathrm{on}}+0.5}\ \exp\left(\pm Z_{\alpha/2}\sqrt{\frac{1}{n_{\mathrm{on}}+0.5}+\frac{1}{n_{\mathrm{tot}}-n_{\mathrm{on}}+0.5}}\right) \tag{26}$$

Tang and Ng [61], in commenting on a paper by Graham et al. [62], examine several methods for intervals for the ratio of Poisson means, including several based on binomial methods. They prefer the adjusted Wald logit method also favored by Price and Bonett, citing them as the source.

Barker and Caldwell [63] compare results of eight methods for 95% C.L. intervals, including Bayesian with uniform and Jeffreys prior. They prefer the Wald log linear method but do not mention the adjustment of adding 0.5 (nor do they cite Price and Bonett); they use instead the C-P interval when $\min(n_{\mathrm{on}}, n_{\mathrm{tot}}-n_{\mathrm{on}}) = 0$. They found that this composite set maintains coverage (even though its theoretical justification relies on asymptotic approximation) and generally performs better than other methods which maintain coverage. If some under-coverage is allowed, they favor Bayesian with uniform prior and the Wilson score interval. (Their criteria include length in the metric uniform in ρ.)

Gu et al. [64] compare four general approaches via Monte Carlo simulation, restricting themselves to the one-sided test. They prefer a test based on a variance-stabilizing transformation of Huffman [65], using an idea of Anscombe [66]. A likelihood ratio test is most powerful against the alternatives they considered, but as it can undercover they advise caution in its use.

Cousins [60] describes his multi-dimensional Neyman construction to obtain a set of *central* confidence intervals for the ratio of Poisson means, obtaining intervals that are strictly conservative in unconditional coverage, and which are always subsets (proper subsets except for $n_{\text{tot}} = 1$) of Clopper-Pearson-derived intervals. The unconditional coverage is obtained by averaging over conditional under-coverage and over-coverage at different values of n_{tot}. When translated back into confidence intervals for a binomial parameter, these intervals are remarkably similar to mid-P intervals, as discussed below.

In performing coverage tests, there is an issue of what to do if the data obtained has *both* n_{off} and n_{on} equal to zero, so that $n_{\text{tot}} = 0$. As nothing has been learned about the ratio, the only sensible confidence interval is $(0, \infty)$, which always covers the unknown true value. Cousins argues in Ref. [60] that such experiments should be excluded from the coverage calculation, since as a practical matter, “An experimenter who observes neither Poisson process will normally make *no* statement regarding the ratio of their means! Thus, practical confidence intervals should have the property that the requisite coverage is obtained when one considers only those experimenters who do not obtain (0,0).” We still believe this to be the case, but note that the coverage calculations of Refs. [16,61,63] include observed data (0,0).

In the remainder of this paper, we discuss in more detail how the ratio-of-Poisson-means problem allows one to evaluate binomial intervals using a frequentist average over data; we find this to be preferable to averaging over the binomial parameter, which requires a choice of metric (or a choice of Bayesian prior from which a corresponding natural metric can be inferred). We then perform this evaluation for a number of the available sets of intervals, and conclude with observations and recommendations.

5 Frequentist evaluation of the performance of the various recipes

Given n_{tot}, ρ, and a C.L., one can use any of the above recipes to obtain the set of $n_{\mathrm{tot}}+1$ intervals $[t_{\mathrm{low}}, t_{\mathrm{up}}]$ for ρ corresponding to the possible observations. As described and illustrated above, in the frequentist evaluation of these intervals, one calculates the probability $\mathrm{Bi}(n_{\mathrm{on}}|n_{\mathrm{tot}}, \rho)$ of obtaining each interval, and thus the probabilities that $\rho < t_{\mathrm{low}}$, that $\rho \in [t_{\mathrm{low}}, t_{\mathrm{up}}]$ and that $\rho > t_{\mathrm{up}}$. For the ratio of Poisson means problem, one is given $(n_{\mathrm{on}}, n_{\mathrm{off}})$, and a C.L., from which $n_{\mathrm{tot}} = n_{\mathrm{on}} + n_{\mathrm{off}}$ and hence an interval for ρ is calculated and then translated into an interval for λ using Eqn. 8. For any given μ_{tot} and λ, probabilities of obtaining $(n_{\mathrm{on}}, n_{\mathrm{off}})$ are calculated from Eqn. 12, and thus unconditional probabilities for covering λ can be calculated from the obtained interval sets and these probabilities.

In Figs. 6 through 19, we plot the probability that the parameter is in the interval for methods which are among those advocated in the above references: Clopper-Pearson with mid-P modification, at 68.27% C.L. (Figs. 6a and b, 17a) and at 95% C.L. (Figs. 7a and b); Wilson score at 68.27% C.L. (Figs. 8a and b); generalized Agresti-Coull at 68.27% C.L. (Figs. 9a and b); Wald log-linear at 68.27% C.L. (Figs. 10a and b); $\Delta(-2\ln\mathcal{L})$ at 68.27% C.L. (Figs. 11a and b); exact LR test inversion, i.e., Neyman construction with likelihood-ratio ordering, with and without mid-P modification, at both C.L.'s (Figs. 12 through 16); Cousins's [60] ratio-of-Poisson means translated into binomial at 68.27% C.L. (Figs. 17b, 18a and b); and Bayesian with prior uniform in ρ at 68.27% C.L. (Figs. 19a and b). Additional plots for nearly all methods mentioned in this paper, for a variety of confidence levels, slices, as well as for one-sided probabilities, are available on request from the authors.

A striking aspect of the two-dimensional plots is the variation of coverage, which is difficult to capture in tables of average coverage or rms of coverage: superimposed on the oscillations are evident trends which indicate regions of particularly low or high coverage. A number of methods give large undercoverage either at low n_{tot} or at ρ near the endpoints; as these values naturally arise in HEP applications, we disfavor such methods.

Another significant observation is how well the mid-P methods perform in the unconditional coverage calculations for the ratio of Poisson means. In effect the Poisson fluctuations of n_{tot} are performing some randomization on top of the "mid" value (0.5) which was fixed in the mid-P calculation for fixed n_{tot}. For central intervals, the result is a remarkable resemblance to the corresponding plots for the central intervals of Cousins [60], which are strictly conservative for the ratio of Poisson means, but much less so than Clopper-Pearson intervals. This similarity was discovered while performing the calculations for this paper, and is seen in Figs. 17a and b; Figs. 6a and 18a; Figs. 6b and 18b; and in

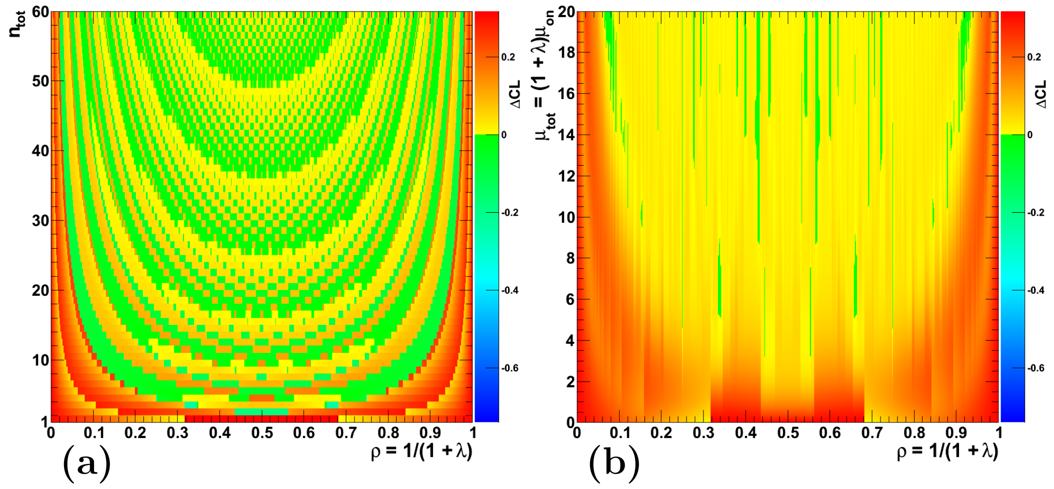

Fig. 6. (a) Coverage of 68.27% C.L. (Clopper-Pearson) mid-P intervals, as a function of ρ and $n_{\rm tot}$, and (b) unconditional coverage of the same intervals for λ. A horizontal slice of (b) is in Fig. 17a.

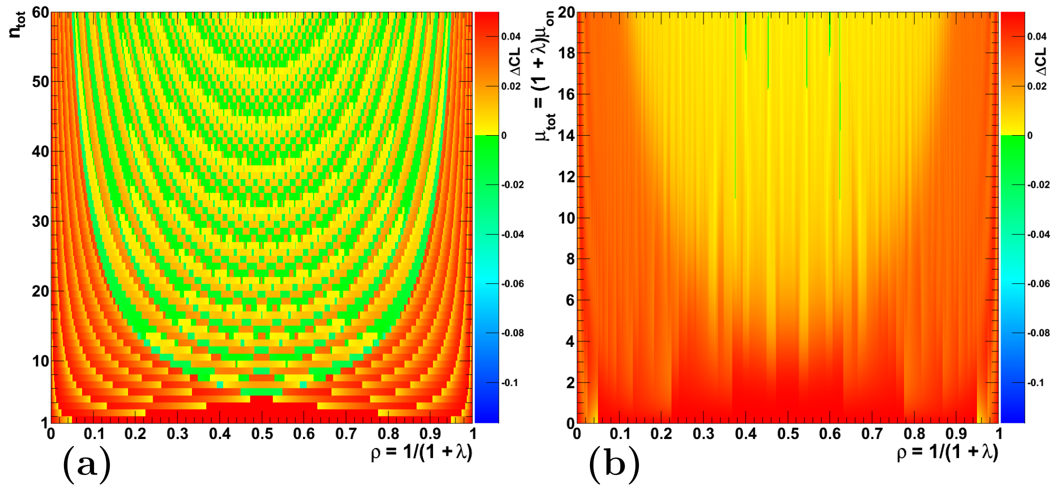

Fig. 7. (a) Coverage of 95% C.L. (Clopper-Pearson) mid-P intervals, as a function of ρ and $n_{\rm tot}$, and (b) unconditional coverage of the same intervals for λ.

numerous other plots inspected by the authors. One could imagine further tuning (as a function of $n_{\rm tot}$) the "mid" value of 0.5 in order to optimize coverage, but we did not explore this.

The Bayesian-inspired methods perform reasonably well with the ad-hoc choice of using central intervals unless $n_{\rm on}$ is 0 or $n_{\rm tot}$, in which case the interval was pushed against the endpoint, as described above. This leads to over-coverage near the endpoints (a feature of many methods). We did not explore alternatives such as highest-posterior-density intervals.

Among asymptotic methods, the Wilson score interval and the (preferred)

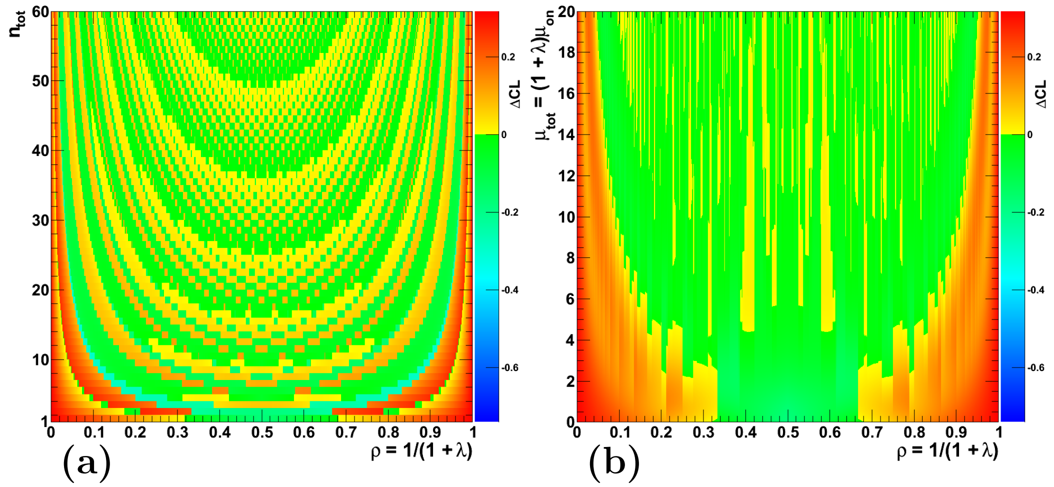

Fig. 8. (a) Coverage of 68.27% C.L. Wilson score intervals, as a function of ρ and $n_{\rm tot}$, and (b) unconditional coverage of the same intervals for λ.

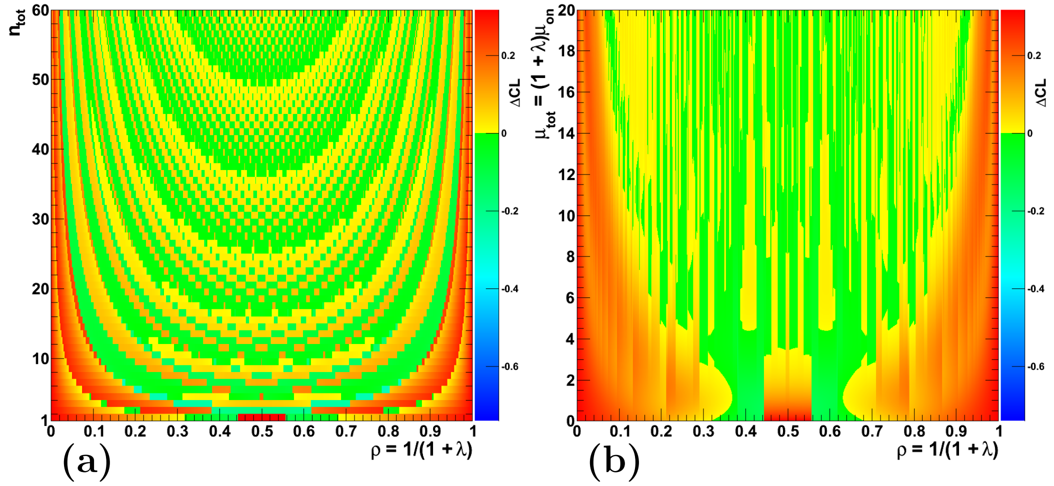

Fig. 9. (a) Coverage of 68.27% C.L. generalized Agresti-Coull intervals, as a function of ρ and $n_{\rm tot}$, and (b) unconditional coverage of the same intervals for λ.

generalized Agresti-Coull interval appear to be reasonable for quick estimates as various authors have advocated. The $\Delta(-2\ln\mathcal{L})$ method undercovers at low $n_{\rm tot}$, and is generally not advocated in the literature reviewed.

6 Conclusion

While intervals such as the Wilson score and the generalized Agresti-Coull can be useful for hand calculations and quick estimates (and are a dramatic improvement over the Wald intervals), the methods based on "exact" calcula-

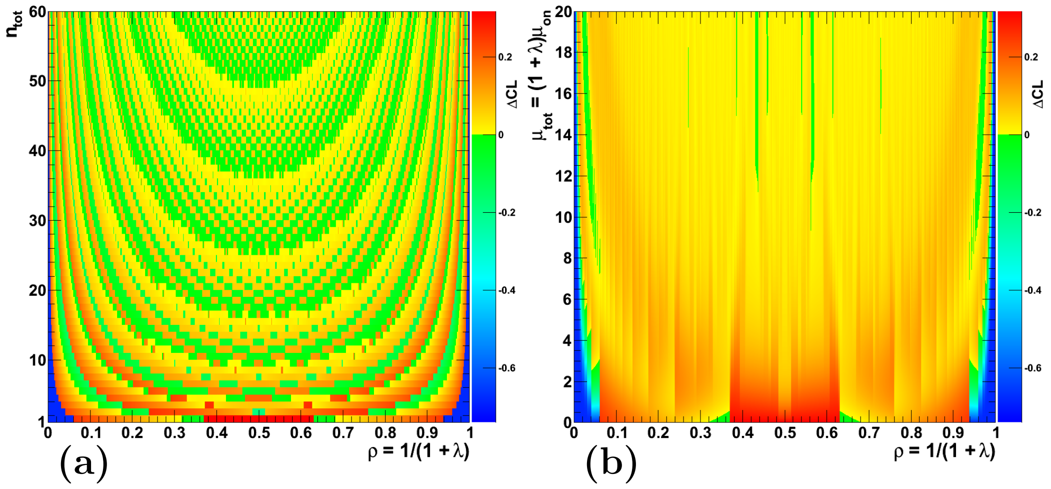

Fig. 10. (a) Coverage of 68.27% C.L. Wald-log-linear intervals, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ.

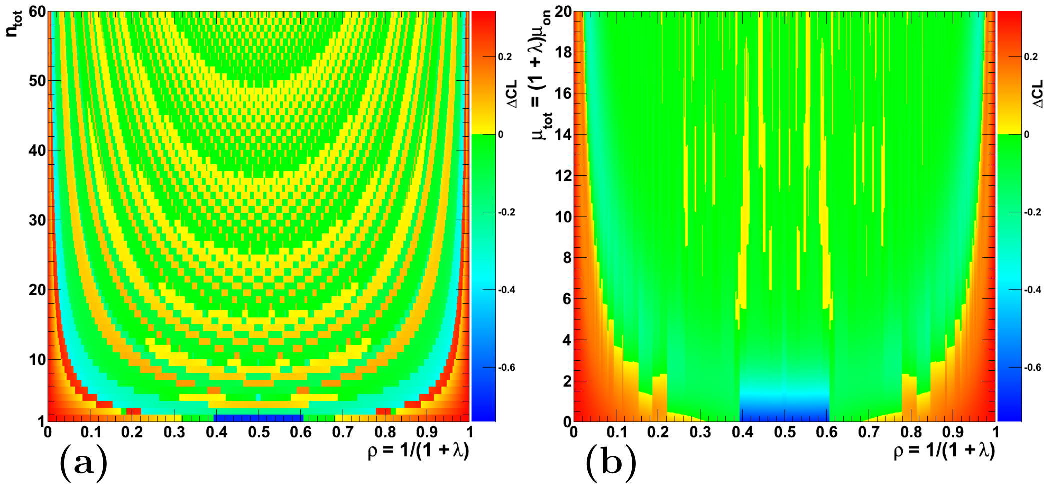

Fig. 11. Coverage of 68.27% C.L. $\Delta(-2\ln\mathcal{L})$ intervals, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ.

tions (i.e., using the binomial and Poisson probabilities rather than asymptotic or Bayesian-inspired calculations) appear to give the most reliable frequentist coverage. When strictly conservative coverage is desired, this statement is a tautology, but it also appears to be the case when approximate coverage is desired, if (as we advocate) the average coverage is evaluated by averaging over data in the closely related ratio-of-Poisson-means problem, rather than attempting to average over ρ.

For *central* intervals, the original *Clopper-Pearson* intervals [1] remain the strictly conservative standard [18], at the cost of severe over-coverage, especially at small n_{tot}. Among the many variants of strictly conservative *both-tailed* (non-central) intervals, we prefer those based on *likelihood-ratio-ordering*,

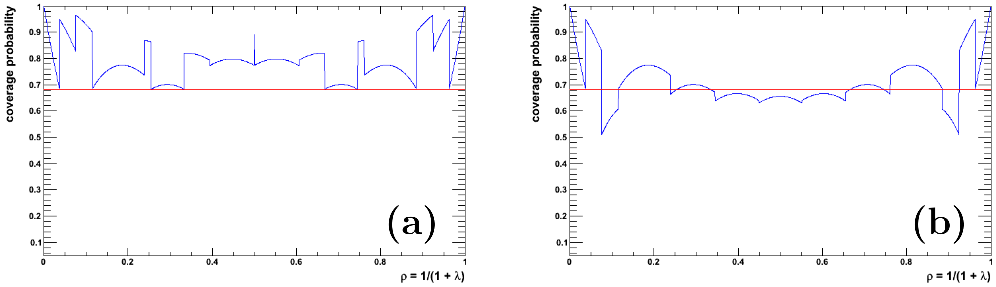

Fig. 12. (a) Coverage of 68.27% C.L. intervals obtained from exact inversion of the LR test, as a function of ρ, for fixed $n_{\text{tot}} = 10$, and (b) coverage of same intervals but with mid-P modification. (a) and (b) are horizontal slices of Figs. 13a and 15a, respectively.

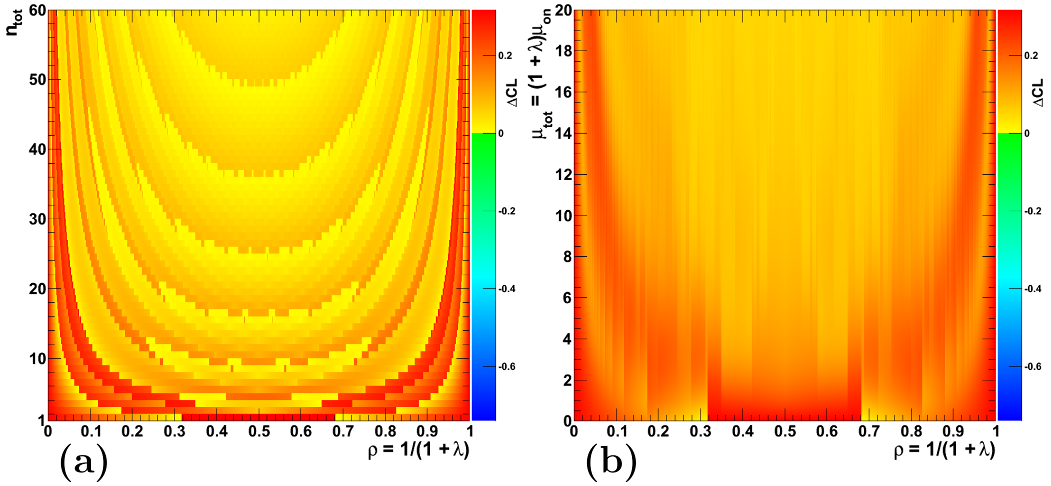

Fig. 13. (a) Coverage of 68.27% C.L. intervals obtained from exact inversion of the LR test, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ. A horizontal slice of (a) is in Fig. 12a.

i.e., the intervals obtained by "exact inversion of the LR test", the method advocated in HEP by Feldman and Cousins [7]. The likelihood-ratio test [6] generalizes well to many complex, multi-dimensional problems in statistical inference [7], and thus is well-integrated into a larger picture; when using more specialized ad hoc manipulations applied to the binomial problem, one is faced with the problem of when to abandon them (and what to replace them with) as more complications are added to the original simple $(n_{\text{on}}, n_{\text{tot}})$ problem.

In the ratio-of-Poisson-means problem, we prefer making *Lancaster's mid-P modification* [32] to the construction of either set of exact intervals in the above paragraph. It provides remarkably good approximate coverage in the ratio-of-Poisson-means problem when evaluated in the unconditional ensemble (i.e., frequentist averaging over values of n_{tot} other than the value observed, weighted by their Poisson probabilities). The mid-P intervals are strikingly

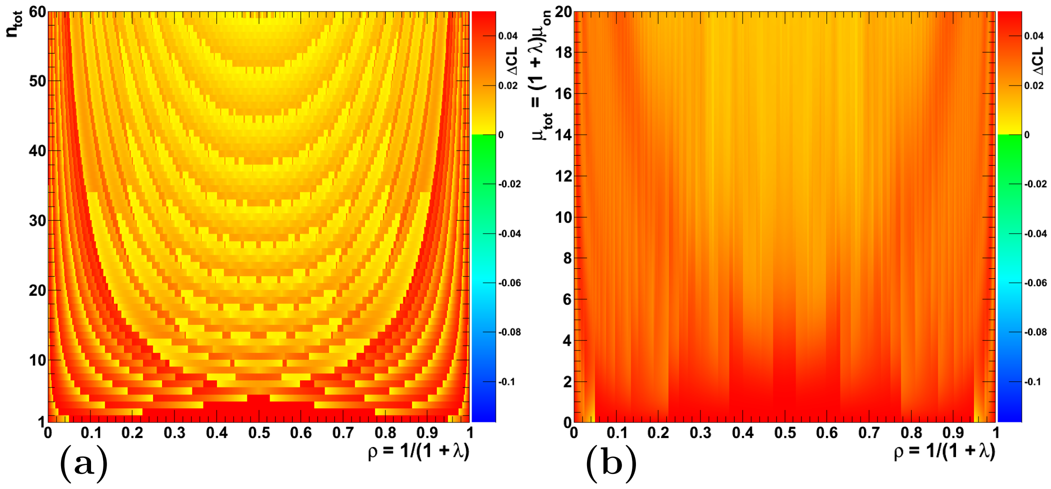

Fig. 14. (a) Coverage of 95% C.L. intervals obtained from exact inversion of the LR test, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ.

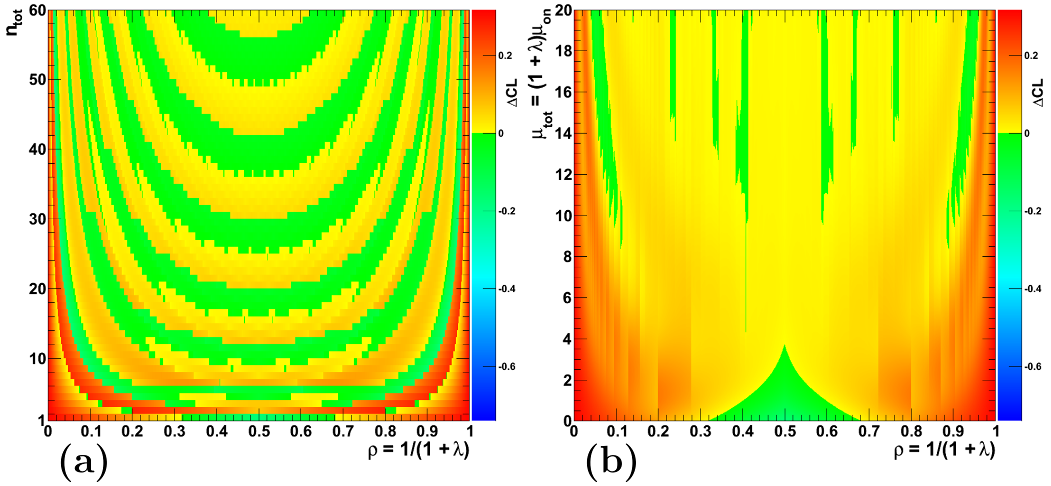

Fig. 15. (a) Coverage of 68.27% C.L. intervals obtained from exact inversion of the LR test with mid-P modification, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ. A horizontal slice of (a) is in Fig. 12b.

similar to a set constructed by Cousins which strictly covers the ratio, but the mid-P intervals have a much simpler description that can also be generalized as complexity is added to the problem. One can also imagine contexts (such as estimating efficiencies of many similar detector elements) in which Poisson fluctuations of the number of trials in each detector element provides a sort of frequentist ensemble which would suggest that mid-P intervals should be considered. However, use of mid-P intervals in a context in which there is no such frequentist averaging would go against the traditional conventions of HEP. Introduction of nuisance parameters (e.g., some systematic uncertainties) into the pure binomial problem, as is common in HEP, can provide another source

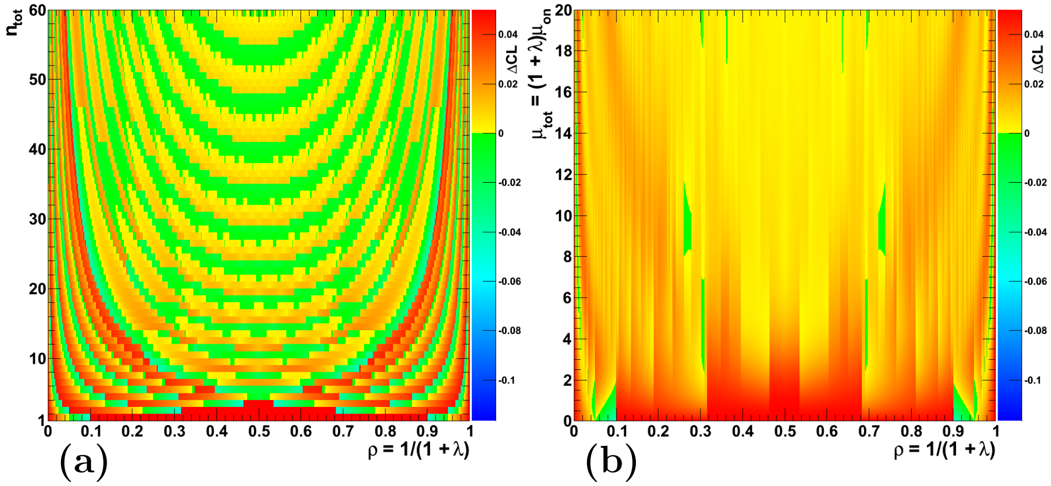

Fig. 16. (a) Coverage of 95% C.L. intervals obtained from exact inversion of the LR test with mid-P modification, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ.

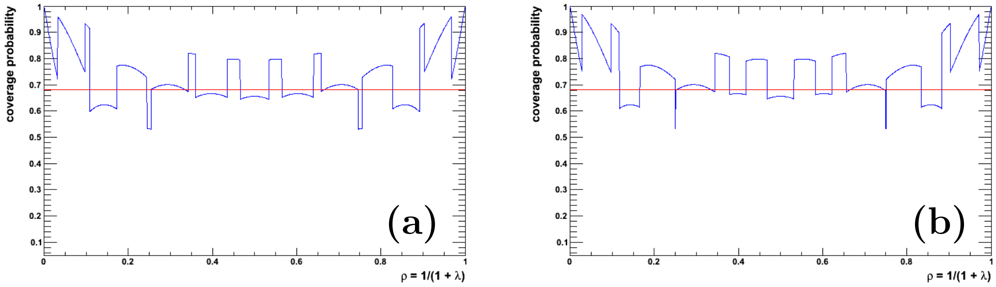

Fig. 17. (a) Coverage of 68.27% C.L. (Clopper-Pearson) mid-P intervals, and (b) coverage of 68.27% C.L. intervals constructed by Cousins [60] for the ratio of Poisson means and translated here to intervals for ρ, both as a function of ρ, for fixed $n_{\mathrm{tot}} = 10$. (a) and (b) are horizontal slices of Figs. 6a and 18a, respectively. The remarkable resemblance is typical of that for other values of ρ and n_{tot}.

of averaging. We speculate that mid-P intervals could prove to be useful for obtaining good coverage in many such contexts.

The use of these intervals can of course be considered in any application of binomial intervals. In high energy and astroparticle physics, the "on/off" (signal bin plus sideband) problem was recently explored in detail by Cousins, Linnemann, and Tucker [67]; one of the promising methods for computing the statistical significance of a signal (denoted by Z_{Bi}) used the Clopper-Pearson interval. In some contexts it should be useful to consider as well one or more of the other three intervals recommended here when calculating Z_{Bi}.

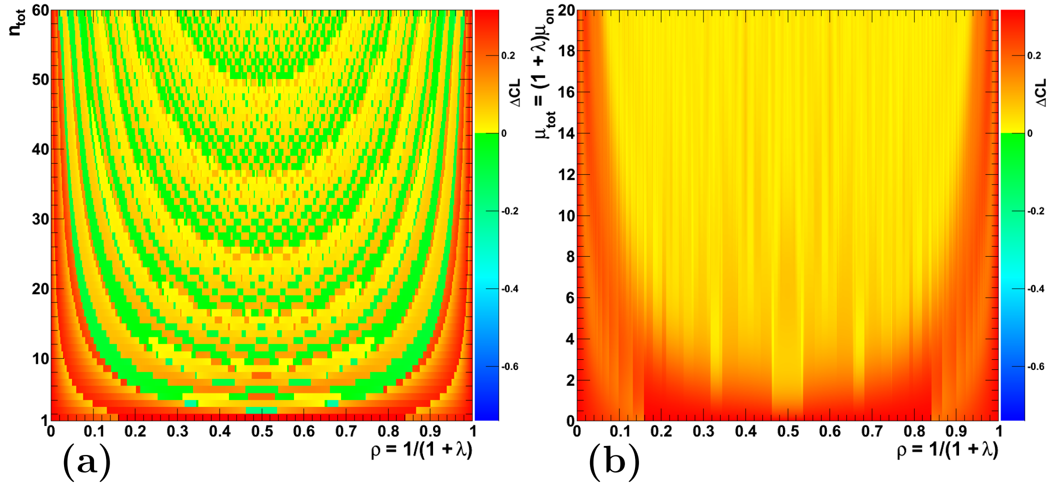

Fig. 18. (a) Coverage of 68.27% C.L. intervals constructed by Cousins [60] for the ratio of Poisson means and translated here to intervals for ρ, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ. A horizontal slice of (a) is in Fig. 17b.

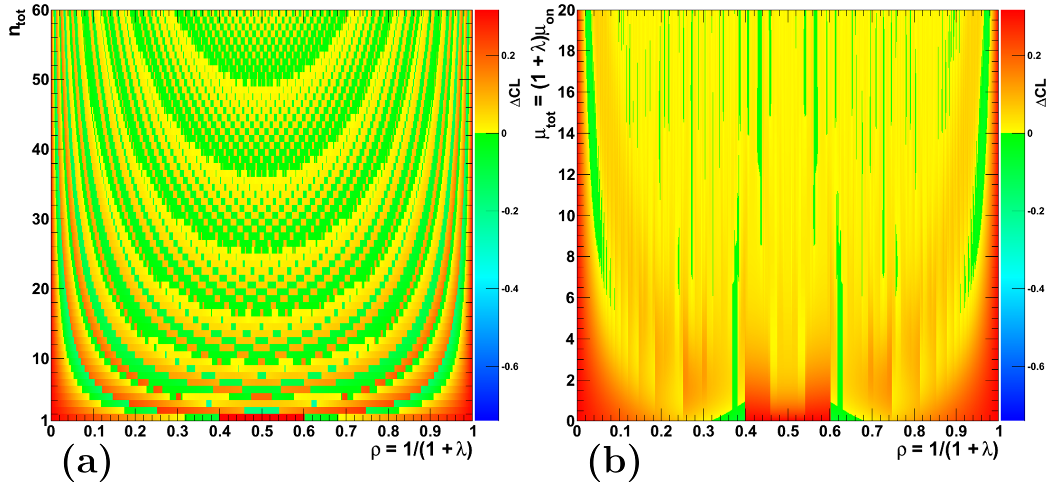

Fig. 19. (a) Coverage of intervals calculated using a Bayesian method with uniform prior and containing 68.27% posterior probability, as a function of ρ and n_{tot}, and (b) unconditional coverage of the same intervals for λ.

Acknowledgements

This work was partially supported by the U.S. Department of Energy and the National Science Foundation.

References

[1] C.J. Clopper and E.S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika **26** (1934) 404.

[2] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta, "Interval Estimation for a Binomial Proportion," Statistical Science **16** (2001) 101. `http://www.jstor.org/stable/2676784`.

[3] J. Przyborowski and H. Wilenski, "Homogeneity of Results in Testing Samples from Poisson Series," Biometrika **31** (1940) 313.

[4] F. James and M. Roos, "Errors on Ratios of Small Numbers of Events," Nuclear Physics **B172** (1980) 475.

[5] N. Reid, "The Roles of Conditioning in Inference," Stat. Sci. **10** (1995) 138.

[6] For a comprehensive discussion of hypothesis testing, see A. Stuart, K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Volume 2A, 6th ed., (London:Arnold, 1999), and earlier editions by Kendall and Stuart. The formal correspondence between hypothesis tests and confidence intervals is discussed in Chapter 20.

[7] Gary J. Feldman and Robert D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. **D57** (1998) 3873.

[8] T. Tony Cai, "One-Sided Confidence Intervals in Discrete Distributions," J. Statistical Planning and Inference **131** (2005) 63.
DOI: 10.1016/j.jspi.2004.01.005.

[9] Edwin B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," J. Amer. Stat. Assoc. **22** (1927) 209. `http://www.jstor.org/stable/2276774`.

[10] Alan Agresti and Brent A. Coull, "Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions," American Statistician **52** (1998) 119. `http://www.jstor.org/stable/2685469`.

[11] Alan Agresti and Brent A. Coull, "[Interval Estimation for a Binomial Proportion]: Comment," Statistical Science **16** (2001) 117. `http://www.jstor.org/stable/2676785`.

[12] S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," Annals of Math. Stat. **9** (1938) 60.

[13] F. James, "Interpretation of the shape of the likelihood function around its minimum," Computer Physics Communications **20** (1980) 29.

[14] D.R. Cox and D.V. Hinkley, *Theoretical Statistics*, Chapman and Hall (1974), reprinted by CRC Press, Boca Raton, FL (2000), pp. 27-28.

[15] John J. Gart, "Alternative Analyses of Contingency Tables," J. Royal Stat. Soc. Series B (Methodological) **28** (1966) 164. http://www.jstor.org/stable/2984283.

[16] Robert M. Price and Douglas G. Bonett, "Estimating the ratio of two Poisson rates," Computational Statistics & Data Analysis **34** (2000) 345.

[17] J. Neyman, Phil. Trans. Royal Soc. London, Series A, **236** 333-80 (1937). Reprinted in *A Selection of Early Statistical Papers on J. Neyman* (University of California Press, Berkeley, 1967), pp. 250-289. See in particular pp. 250-252, 261-268, 285-286, of the reprint. The quantity which Neyman calls α is $(1-\alpha)$ in most modern references.

[18] C. Amsler et al. (Particle Data Group), "Review of Particle Physics," Physics Letters **B667** (2008) 1. An expression for the endpoints of the Clopper-Pearson intervals (in terms of the F distribution) has been in the PDG Review of Particle Physics since 2002.

[19] John E. Angus and Ray E. Schafer, "Improved Confidence Statements for the Binomial Parameter," American Statistician **38** (1984) 189, http://www.jstor.org/stable/2683650.

[20] Theodore E. Sterne, "Some Remarks on Confidence of Fiducial Limits," Biometrika **41** (1954) 275, http://www.jstor.org/stable/2333026.

[21] Edwin R. Crow, "Confidence Intervals for a Proportion," Biometrika **43** (1956) 423. doi:10.1093/biomet/43.3-4.423, http://biomet.oxfordjournals.org/cgi/content/citation/43/3-4/423.

[22] Colin R. Blyth and Harold A. Still, "Binomial Confidence Intervals," J. Amer. Stat. Assoc. **78** (1983) 108. http://www.jstor.org/stable/2287116.

[23] George Casella, "Refining Binomial Confidence Intervals," Canadian Journal of Statistics / La Revue Canadienne de Statistique **14** (1986) 113. http://www.jstor.org/stable/3314658.

[24] George Casella, "[Interval Estimation for a Binomial Proportion]: Comment," Statistical Science **16** (2001) 120. http://www.jstor.org/stable/2676786.

[25] Helge Blaker, "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions," Canadian J. Statistics / La Revue Canadienne de Statistique **28** (2000) 783, http://www.jstor.org/stable/3315916. See also corrections in "Corrigenda," Vol. **29** (2001) 681, http://www.jstor.org/stable/3316015.

[26] Helga Blaker and Emil Spjotvoll, "Paradoxes and Improvements in Interval Estimation," American Statistician **54** (2000) 242, http://www.jstor.org/stable/2685774.

[27] Paul W. Vos and Suzanne Hudson, "Problems with binomial two-sided tests and the associated confidence intervals," Australian & New Zealand J. Statistics **50** (2008) 81, DOI: 10.1111/j.1467-842X.2007.00501.x.

[28] Chris Corcoran and Cyrus Mehta, "[Interval Estimation for a Binomial Proportion]: Comment," Statistical Science **16** (2001) 122. http://www.jstor.org/stable/2676787.

[29] Gioacchino Ranucci, "Binomial and ratio-of-Poisson-means frequentist confidence intervals applied to the error evaluation of cut efficiencies," arXiv:0901.4845v1 [physics.data-an].

[30] W.L. Stevens, "Fiducial limits of the parameter of a discontinuous distribution," Biometrika 37 (1950) 117; Biometrika 44 (1957) 436.

[31] Robert Cousins, "A method which eliminates the discreteness in Poisson confidence limits and lessens the effect of moving cuts specifically to eliminate candidate events," Nucl. Instrum. Meth. A **337**, 557 (1994).

[32] H.O. Lancaster, "Significance Tests in Discrete Distributions," J. Amer. Stat. Assoc. **56** (1961) 223. http://www.jstor.org/stable/2282247.

[33] G. Berry and P. Armitage, "Mid-P Confidence Intervals: A Brief Review," The Statistician **44** (1995) 417. http://www.jstor.org/stable/2348891.

[34] Alan Agresti and Anna Gottard, "Comment:Randomized Confidence Intervals and the Mid-P Approach," Statistical Science, **20** (2005) 367. http://www.jstor.org/stable/20061194, DOI: 10.1214/088342305000000403.

[35] Alan Agresti and Anna Gottard, "Nonconservative exact small-sample inference for discrete data," Computational Statistics & Data Analysis **51** (2007) 6447. DOI:10.1016/j.csda.2007.02.024.

[36] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta, "Comment: Fuzzy and Randomized Confidence Intervals and P-Values," Statistical Science **20** (2005) 375. DOI: 10.1214/088342305000000395.

[37] Anthony O'Hagan, *Kendall's Advanced Theory of Statistics*, Volume 2B Bayesian Inference, 1st ed., (London:Arnold, 1994). The index points to several illustrative examples using the binomial distribution.

[38] N. Reid, R. Mukerjee, D. A. S. Fraser, "Some aspects of matching priors," in Mathematical statistics and applications: Festschrift for Constance van Eeden (Beachwood, OH: Institute of Mathematical Statistics, 2003), 31-43. ed. by Marc Moore, Sorana Froda, and Christian Léger. http://projecteuclid.org/euclid.lnms/1215091929.

[39] P.D. Baines and X.-L. Meng, "Probability Matching Priors in LHC Physics," in Proceedings of the PHYSTAT LHC Workshop on Statistical Issues for LHC Physics (CERN, Geneva, 27-29 June 2007), CERN Yellow Report 2008-001, ed. by L. Lyons, H. Prosper, and A. de Roeck. The talk contains many more details: see http://phystat-lhc.web.cern.ch/phystat-lhc/program.html.

[40] B. L. Welch and H. W. Peers, "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," J. Royal Stat. Soc. Series B (Methodological) **25** (1963) 318. http://www.jstor.org/stable/2984298.

[41] B.L. Welch, "On Comparisons Between Confidence Point Procedures in the Case of a Single Parameter," J. Royal Stat. Soc. Series B (Methodological) **27** (1965) 1. http://www.jstor.org/stable/2984475.

[42] H.W. Peers, "On Confidence Points and Bayesian Probability Points in the Case of Several Parameters," J. Royal Stat. Soc. Series B (Methodological) **27** (1965) 9. http://www.jstor.org/stable/2984476.

[43] Harold Jeffreys, *Theory of Probability* (Oxford University Press, New York, 1961), 3rd ed.

[44] Robert E. Kass and Larry Wasserman, "The Selection of Prior Distributions by Formal Rules," J. Am. Stat. Assoc. **91** (1996) 1343. http://www.jstor.org/stable/2291752, http://lib.stat.cmu.edu/~kass/papers/rules.pdf.

[45] James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed., (New York: Springer-Verlag, 1980).

[46] Seymour Geisser, "On Prior Distributions for Binary Trials," American Statistician **38** (1984) 244.
http://www.jstor.org/stable/2683393; including Comments by Jose M. Bernardo, http://www.jstor.org/stable/2683394; by Melvin R. Novick http://www.jstor.org/stable/2683395; by Arnold Zellner http://www.jstor.org/stable/2683396; and reply by Seymour Geisser http://www.jstor.org/stable/2683397.

[47] D.J. Brenner and H. Quan, "Exact Confidence Limits for Binomial Proportions–Pearson & Hartley Revisited," The Statistician **39** (1990) 391, http://www.jstor.org/stable/2349083.

[48] J. B. Copas, "Exact Confidence Limits for Binomial Proportions–Brenner & Quan Revisited," The Statistician **41** (1992) 569, http://www.jstor.org/stable/2348922.

[49] Donald B. Rubin and Nathaniel Schenker, "Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior," Sociological Methodology, **17** (1987) 131. http://www.jstor.org/stable/271031.

[50] Jenõ Reiczigel, "Confidence intervals for the binomial parameter: some new considerations," Statistics in Medicine **22** (2003) 611, DOI: 10.1002/sim.1320.

[51] A. Agresti, "Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact," Stat Methods Med Res **12** (2003) 3. http://smm.sagepub.com/cgi/content/abstract/12/1/3, DOI: 10.1191/0962280203sm311ra.

[52] Borek Puza and Terence O'Neill, "Generalised Clopper-Pearson confidence intervals for the binomial proportion," J. of Statistical Computation & Simulation **76** (2006) 489, DOI:10.1080/10629360500107527.

[53] Stein Emil Vollset, "Confidence Intervals for a Binomial Proportion," Statistics in Medicine **12** (1993) 809.

[54] Michael D. deB. Edwardes, "The evaluation of confidence sets with application to binomial intervals," Statistica Sinica **8** (1998) 393,' `http://www3.stat.sinica.edu.tw/statistica/`.

[55] Robert G. Newcombe, "Two-sided confidence intervals for the single proportion: comparison of seven methods," Statistics in Medicine **17** (1998) 857; and Comment by Steven A. Julious, Statistics in Medicine **24** (2005) 3383 regarding numerical evaluation.

[56] Wang-Shu Lu, "Improved confidence intervals for a binomial parameter using the Bayesian method," Communications in Statistics - Theory and Methods **29** (2000) 2835, `http://www.informaworld.com/10.1080/03610920008832639`.

[57] Colin R. Blyth, "Approximate Binomial Confidence Limits," J. Amer. Stat. Assoc. **81** (1986) 843. `http://www.jstor.org/stable/2289018`.

[58] Alan Agresti and Yongyi Min, "On Small-Sample Confidence Intervals for Parameters in Discrete Distributions," Biometrics **57** (2001) 963. `http://www.jstor.org/stable/3068439`.

[59] Ana M. Pires and Conceição Amado, "Interval estimators for a binomial proportion: comparison of twenty methods," REVSTAT – Statistical Journal **6** (2008) 165. `http://www.ine.pt/revstat/pdf/rs080204.pdf`.

[60] R. D. Cousins, "Improved central confidence intervals for the ratio of Poisson means," Nucl. Instrum. Meth. A **417** (1998) 391.

[61] Man-Lai Tang and Hon Keung Tony Ng, "Comment on: Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method," Statistics in Medicine **23** (2004) 685, DOI: 10.1002/sim.1405.1683; with reply DOI: 10.1002/sim.1405.1684.

[62] P.L. Graham, K. Mengerson, and A.P. Morton, "Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method," Statistics in Medicine **22** (2003) 2071, DOI: 10.1002/sim.1405.

[63] Lawrence Barker and Betsy L. Cadwell, "An analysis of eight 95 per cent confidence intervals for a ratio of Poisson parameters when events are rare," Statistics in Medicine **27** (2008) 4030, DOI: 10.1002/sim.3234 `http://dx.doi.org/10.1002/sim.3234`.

[64] Kangzia Gu, Hon Keung Tony Ng, Man Lai Tang, and William R. Schucany, "Testing the ratio of Two Poisson Rates," Biometrical Journal **50** (2008) 283. DOI: 10.1002/bimj.200710403.

[65] Michael D. Huffman, "An Improved Approximate Two-sample Poisson Test," Applied Statistics **33** (1984) 224. `http://www.jstor.org/stable/2347448`.

[66] F.J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data," Biometrika **35** (1948) 246. DOI: 10.1093/biomet/35.3-4.246.

[67] Robert D. Cousins, James T. Linnemann, Jordan Tucker, “Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process,” Nucl. Instrum. Meth. A **595** (2008) 480. DOI:10.1016/j.nima.2008.07.086, arXiv:physics/0702156.